CONTROLLED CAPACITY MODELING TOOL

RELATED APPLICATIONS

[0001] This is a continuation-in-part of U.S. Serial No. 09/858,814, filed May 16, 2001, which is a continuation of International Application No. PCT/FR99/02810, with an international filing date of November 16, 1999, which is based on French Patent Application No. 98/14422, filed November 17, 1998.

FIELD OF THE INVENTION

[0002] This invention relates to machine learning and modeling methods.

BACKGROUND

[0003] Modeling methods are known generally in the state of the art, for example from the following articles:

- HOSKING J.M.R. et al. "A statistical perspective on data mining" FUTURE GENERATION COMPUTER SYSTEMS, Nov. 1997, ELSEVIER, Vol. 13, No. 2-3, pages 117-134, which describes the use of statistical methods for processing large volumes of data;

- ROSSIUS R. et al. "A short note about the application of polynomial kernels with fractional degree in support vector learning" MACHINE LEARNING ECML-98. 10th European Conference on Machine Learning. Proceedings CHEMNITZ April 21-23, 1998, pages 143-148, pertaining to the application of classification and forecasting methods.

[0004] The learning problem can be considered as a problem of finding dependencies using a limited number of observations. Thus, it is a question of choosing from a given set of functions $f(x,\alpha)$, $\alpha \in A$, where A is a set of parameters, the one which best approximates the output.

[0005] If $L(y,f(x,\alpha))$ is a measure of the deviation between the real output y and the output predicted by the model $f(x,\alpha)$, it is thus necessary to minimize the effective risk:

$$R(\alpha) = \int L(y, f(x,\alpha)) \, dF(x,y) \qquad \text{(Eq. 1)}$$

while knowing that the joint probability distribution function $F(x,y)$ is unknown and that the only available information is contained in the k observations $(x_1, y_1)...,(x_k, y_k)$ from the learning set.

[0006]   Classically, one determines the function that minimizes the empirical risk calculated on the basis of the learning set:

$$Remp = \sum_{i=1}^{k} L(y_i, f(x_i, \alpha)) \qquad \text{(Eq. 2)}$$

[0007]   One then postulates that this function would be the best approximation of the function that minimizes the effective risk given by (Eq. 1).

[0008]   The problem posed is to know the extent to which a system constructed on the empirical risk minimization inductive principle (Eq. 2) is generalizable, i.e., enables minimizing the effective risk (Eq. 1) including data that have not been learned.

[0009]   Mathematically, a problem is said to be well-posed when it allows a unique solution and this solution is stable, i.e., that a small deviation in the initial conditions can only modify in an infinitesimal manner the form of the solutions. Problems that do not satisfy these conditions are referred to as ill-posed problems.

[0010]   It occurs frequently that the problem of finding f satisfying the equality $A.f=u$ is ill-posed: even if there exists a unique solution to this equation, a small deviation of the right-hand side of this equation can cause large deviations in the solution.

[0011]   And thus if the right-hand member is not exact (u_ε instead of u with $\|u - u_\varepsilon\| \leq \varepsilon$), the functions that minimize the empirical risk $R(f)=\|Af - u_\varepsilon\|^2$ are not necessarily good approximations of the desired solution, even if ε tends to 0.

[0012]   An improvement in solving such problems consists in minimizing another so-called regularized functional of the form:

$$R(f) = R(f) + \lambda (\varepsilon)\Omega (f) \qquad \text{(Eq. 3)}$$

where:

$-\Omega (f)$ is some functional belonging to a special type of operators referred to as regularizing;

$- \lambda(\epsilon)$ is an appropriately chosen constant depending on the level of noise existing on the data.

[0013]One then obtains a sequence of solutions that converge to the desired one as ϵ tends to 0. Minimizing the regularized risk rather than the empirical risk allows obtaining from a limited number of observations a solution that is generalizable to any case.

[0014]Introduction of the regularizing term makes it feasible to provide with certainty a unique solution to an ill-posed problem. This solution can be slightly less accurate than the classic one, but it possesses the fundamental property of being stable, thus endowing the results with greater robustness.

[0015]The methods for solving ill-posed problems demonstrate that there exist other inductive principles that enable obtaining a better regularization capacity than the principle consisting in minimizing the error made on the learning set.

[0016]Therefore, the main objective of theoretical analysis is to find the principles making it feasible to control the generalization capacity of learning systems and to construct algorithms that implement these principles.

[0017]Vapnik's theory is the tool that establishes the necessary and sufficient conditions for a learning process based on the empirical risk minimization principle to be generalizable, leading to a new inductive principle referred to as the structural risk minimization principle. It can be demonstrated that the effective risk satisfies an inequality of the form:

$$R (\alpha)<R_{emp} (\alpha)+F(h,k) \qquad\qquad (\text{Eq. 4})$$

where:

$- h$ is the Vapnik-Chervonenkis dimension of the set of functions $f(x,\alpha)$ among which

the solution is sought;

     - k is the number of observations available for constructing the model;

     - F is an increasing function of h and a decreasing function of k.

[0018]It can be seen immediately that, since the number k of available observations is finite, the fact of minimizing the empirical error is not sufficient for minimizing the effective error. The general idea of the structural risk minimization principle is to take into account the two terms of the right-hand member of the equation (Eq. 4), rather than only the empirical risk. This implies constraining the structure of the set of the functions $f(x, \alpha)$ among which the solution is sought so as to limit or even control the parameter h.

[0019] According to this principle, the development of new algorithms enabling control of the robustness of learning processes would be advantageous.

SUMMARY OF THE INVENTION

[0020] This invention relates to a process for modeling numerical data from a data set including collecting data for development of a model with a data acquisition module, processing the data to enhance its exploitability in a data preparation module, constructing a model by learning on the processed data in a modeling module, evaluating the fit and robustness of the obtained model in a performance analysis module, adjusting the model parameters to select the optimal model in an optimization module, wherein the model is generated in the form of a Dth order polynomial of the variables used in input of the modeling module, by controlling the trade-off between the learning accuracy and the learning stability with the addition to the covariance matrix of a perturbation during calculation of the model in the form of the product of a scalar λ times a matrix H or in the form of a matrix H dependent on a vector of k parameters $\Lambda = (\lambda_1, \lambda_2, \dots \lambda_k)$ where the order D of the polynomial and the scalar λ, or the vector of parameters Λ, are determined automatically during model adjustment by the optimization module by integrating an additional data partition step performed by a partition module which consists in constructing two preferably disjoint subsets: a first subset comprising training data used as a learning

-4-

base for the modeling module and a second subset comprising generalization data destined to adjust the value of these parameters according to a model validity criterion obtained on data that did not participate in the training, and where the matrix H is a positive defined matrix of dimensions equal to the number p of input variables into the modeling module, plus one.

[0021] The invention also relates to a device for modeling numerical data from a data sample including means for acquiring input data, means for processing the input data, means for constructing a model by learning on processed data, means for analyzing performances of the obtained model, means for optimizing the obtained model, wherein the model is generated in the form of a Dth order polynomial of the variables used in input of the modeling module, by controlling the trade-off between the learning accuracy and the learning stability with the addition to the covariance matrix of a perturbation during calculation of the model in the form of the product of a scalar λ times a matrix H or in the form of a matrix H dependent on a vector of k parameters $\Lambda = (\lambda_1, \lambda_2, \ldots \lambda_k)$ where the order D of the polynomial and the scalar λ, or the vector of parameters Λ, are determined automatically during model adjustment by the optimization module by integrating additional means for splitting the data so as to construct two preferably disjoint subsets: a first subset comprising training data used as a learning base for the modeling module and a second subset comprising generalization data destined to adjust the value of these parameters according to a model validity criterion obtained on data that did not participate in the training, and where the matrix H is a positive defined matrix of dimensions equal to the number p of input variables into the modeling module, plus one.

BRIEF DESCRIPTION OF THE DRAWING

[0022] The drawing is a flow/block diagram of a modeling device/process in accordance with aspects of the invention.

DETAILED DESCRIPTION

[0023] The invention provides a model for forecasting the evolution of a phenomenon from

a numerical data set of any size. It can be implemented in the form of specifically designed integrated circuits and then present itself in the form of a specific element functioning in an independent manner. It can also be implemented in software form and be integrated in a computer program. It can, especially, be used for processing a digital signal in an electronic circuit. In a more general application, it enables the modeling of nonlinear phenomena, the analysis of phenomena by means of immediately exploitable formulas and the generation of robust models. The accuracy enabled by these novel methods permits an appreciable increase in machine learning rates.

[0024] The invention can also be used in the domain of risk analysis by insurance companies. These companies store, in a form that is structured to varying degrees, the characteristics of drivers, their vehicles and accidents they have been involved in or caused. It is possible to determine which are at high risk based on these available elements.

[0025] In the modeling of physical phenomena, the analyzed events correspond generally to the data captured by the various sensors in the measurement line. It is possible, for example, to determine which are the combinations of factors that are the source of defective products and, thus, anticipate problems and improve productivity.

[0026] In the domain of flow management, these events would correspond instead to information collected over time. It is possible, for example, to determine the relations existing among the considered flows and the calendar data, or variables that are more specific to the application under consideration such as meteorological data for the consumption of electricity or promotional periods for sales analysis, which enables better management of stocks and of workloads of manufacturers.

[0027] In the banking sector, the events would represent, on the one hand, the profile of the clients and, on the other hand, a description of the operations. The modeling would reveal, for example, the risk factors linked to individuals and to operations.

[0028] The invention pertains to a new modeling technology of very general application, the essential characteristics of which concern the efficiency of the method, the simplicity of the

obtained models and their robustness, i.e., their performance on data that have not been used for learning. The implementation of this technique in an electronic or mechanical information-processing system equipped with sensors and with model exploitation functions enables conception of a tool capable of adapting to and controlling an environment in which there exist complex and changing phenomena, and in which the sensors only partially report the set of the phenomena brought into play. Furthermore, the utmost simplicity of the obtained models provides the user of the tool with an intuitive comprehension of the phenomena he seeks to control.

[0029] The invention uses both classical techniques, such as calculation of covariance matrices, as well as more recent theories, such as those of statistical regularization and consistency of learning processes. The invention consists in that the covariance matrices are not used as such but according to a new process which consists on the one hand in perturbing the covariance matrix in a certain way and on the other hand in adjusting the level of added noise in another way. The way to add and control noise to the data will be described here mathematically but it is feasible to implement these operations in an electronic or mechanical way.

[0030] The invention consists in a process for modeling numerical data from a data set, comprising an input data acquisition step which consists in collecting the data required for the development of the models, performed by a data acquisition module, an input data preparation step which consists in processing the data so as to make them more exploitable, performed by a data preparation module, a modeling step which consists in constructing a model by learning on the processed data, performed by a modeling module, a performance analysis step which consists in evaluating the fit and the robustness of the obtained model, performed by a performance analysis module, a model optimization step which consists in adjusting the parameters of the model so as to select the optimal model, performed by an optimization module, characterized in that one generates the model in the form of a Dth order polynomial of the variables used in input of the modeling module, by controlling the trade-

-7-

off between the learning accuracy and the learning stability with the addition to the covariance matrix of a perturbation during calculation of the model in the form of the product of a scalar λ times a matrix H or in the form of a matrix H dependent on a vector of k parameters $\Lambda = (\lambda_1, \lambda_2, \dots \lambda_k)$ where the order D of the polynomial and the scalar λ, or the vector of parameters Λ, are determined automatically during model adjustment by the optimization module by integrating an additional data partition step performed by a partition module which consists in constructing two preferably disjoint subsets: a first subset comprising training data used as a learning base for the modeling module and a second subset comprising generalization data destined to adjust the value of these parameters according to a model validity criterion obtained on data that did not participate in the training, and where the matrix H is a positive defined matrix of dimensions equal to the number p of input variables into the modeling module, plus one.

[0031] The matrix H can be such that H(p+1,p+1) is different from at least one of the terms H(i,i) for i = 1, 2,..., p.

[0032] Subsequently, two numbers are considered to be close when their relative difference is less than about 10%.

[0033] The matrix H advantageously verifies the following conditions: H(i,i) is close to 1 for i = 1, 2,..., p, H(p+1,p+1) is close to 0 and H(i,j) is close to 0 for i different from j. In a variant, the matrix H verifies the following conditions: H(i,i) is close to a variable a for i = 1, 2,..., p, H(p+1,p+1) is close to a variable b, H(i,j) is close to a variable c for i different from j with a = b + c.

[0034] In an advantageous variant, the matrix H verifies the following additional conditions: a is close to 1-1/p, b is close to 1, c is close to -1/p, where p is the number of variables of the model.

[0035] The base data partition can be performed by an external software program of the spreadsheet or database type, or by specific tools.

[0036] To construct the two training and generalization subsets, one advantageously

performs a pseudorandom sampling, or a pseudorandom sampling keeping the representativeness of the input vectors in the two subsets, or a sequential sampling. In a variant, one performs a first partition of the data into a subset containing the training and generalization data and a subset containing test data.

[0037] One preferred variant consists in selecting the pair of parameters (D, λ) or (D, Λ), which minimizes the mean error, or the weighted mean error, or the mean quadratic error or the weighted mean quadratic error on the generalization data subset.

[0038] The data are advantageously pre-processed by statistical normalization of columns of data, by reconstitution of missing data or by detection and possible correction of outlying data.

[0039] This preparation can be performed by a monovariable or multivariable polynomial development applied to all or part of the inputs, by a trigonometric development of the inputs or by an explicative development of dated inputs.

[0040] One preferred variant consists in using a change of coordinates stemming from a principal components analysis with possible simplification or using one or more temporal shifts before or after all or part of the columns containing the time variables.

[0041] One can advantageously add a preparation explorer which uses a description of the possible processings by the user and an exploration strategy based either on a pure performance criterion in training or in generalization, or on a trade-off between these performances and the capacity of the learning process obtained.

[0042] In one variant, one adds to the modeling process an exploitation module providing monovariable or multivariable polynomial formulas descriptive of the phenomenon, trigonometric formulas descriptive of the phenomenon, or descriptive formulas of the phenomenon containing date developments in calendar indicators.

[0043] The general synopsis of the invention is presented in the Drawing. It comprises all or part of the following elements:

- a data acquisition module (1);

- a data preparation module (2);

- a modeling module (3);

- a performance analysis module (4);

- an optimization module (5);

- a preparation exploration module (6);

- an exploitation module (7).

[0044] The purpose of the data acquisition module (1) is to collect information required for the elaboration of the models. The collection is performed by means of acquisition configuration information, which is transmitted by an operator, either once and for all upon the conception of the system, or in a dynamic manner as a function of new requirements identified over the course of its exploitation. The data can be collected by means of sensors of physical measurements, or in databases by means of requests, or both. In configuring the acquisition, the operator defines for the tool a modeling problem to be treated. On demand, this module produces a rough history of the phenomenon, characterized by a table containing in columns the characteristic parameters of the phenomena (stemming for example from sensors) and in rows the events, each of which corresponds to one observation of the phenomenon. This historic table can be supplemented by a description of the data consisting of information that can be useful for the modeling, and then for the exploitation of the models. The description typically includes the following information:

- name of the column;

- reference of the associated sensor;

- nature of the data (boolean, integer, numerical, dated, region, etc.).

[0045] The data preparation module (2), also referred to as the data processing module, enables refinement of the characteristics of the raw data stemming from the acquisition. Based on the historic table and the data description, this module provides a more complex table in which each column is obtained from a processing operating on one or more columns of the historic table. The processes performed on a column can be in particular:

-10-

- a transformation of the column by a classic function (log, exp, sin, etc.), with each element of the column being substituted by the result of the selected function;

- a K order monovariable polynomial development, generating K columns from one input column x, corresponding to the variables $x, x^2, \ldots, x^K$;

- a spectral development of period T and of order N, generating 2K columns from one input column x, the first K columns being equal to $\cos(2\pi ix/T)$ (for i between 1 and K), and the last K columns being equal to $\sin(2\pi ix/T)$ (for i between 1 and K);

- a development in calendar indicators, generating for one dated input column a list of more specific indicators representative of the events associated with this date (annual, monthly, weekly development, Boolean indicators for the day of the week, holiday, extended weekend, day before extended weekend, day after extended weekend, indicators of holiday, of beginning and end of holidays specific to each region, etc.).

[0046] The data preparation module can also act on multiple columns or multiple groups of columns. It can especially perform the following constructions:

- based on a dated column and a region column, the preparator can carry out a development in meteorological indicators (wind, precipitation, hygrometry, etc.) for the day itself or adjacent days. This operation is executed from a meteorological database;

- based on two groups of columns G1 and G2, the preparator can create a new group of columns G3 comprising the cross-products between all of the columns of the two groups;

- based on a group of columns G, comprising p variables $x_1, x_2, \ldots x_p$, the preparator can generate all the polynomial terms of degree less than or equal to K, thus a group of columns each comprising a term of the type $(x_1)^{K1}(x_2)^{K2}\ldots(x_p)^{Kp}$ with $(K1+\ldots+Kp) \leq K$, with all of the Ki being between 0 and K.

[0047] The data preparation module can also execute operations on rows, notably:

- centering, which subtracts from each element of a column the mean obtained on its column;

- reduction, which divides each element of a column by the standard deviation of its

-11-

column;

- statistical normalization which links together the two preceding operations.

[0048] The data preparation module can also execute global operations in a manner especially so as to reduce the dimension of the problem:

- elimination of a column if its standard deviation is zero;

- elimination of a column whose correlation with a preceding column is greater than a threshold;

- elimination of a column whose correlation with the output is less than a threshold;

- realization of a principal components analysis which leads to a change of coordinates by favoring the principal axes of representation of the phenomenon, and the possible elimination of nonsignificant columns.

[0049] The data preparation module also enables defining the processing of missing values. An event (row) containing one or more missing values will be ignored by default. Nevertheless, the user can fill in the missing value according to various criteria:

- mean of the value on the column;

- mean of the value on a subset of the column;

- the most frequent value (Boolean or enumerated);

- selection of a fixed substitution value;

- estimation of this value based on a modeling function of other variables.

[0050] Another manner of processing missing values is to consider them as a particular value of the variable that can be taken into account by creating an additional boolean column indicating whether the value is present or not.

[0051] The data preparation module also enables detection and processing of suspicious values. Detection is based on the following criteria:

- data outside a range defined by the operator;

- data outside a range calculated by the system (for example, range centered on the mean value and large by K times the standard deviation, analysis of the extreme percentiles,

etc.);

- for Boolean or enumerated data, values whose number of occurrences is less than a given threshold.

[0052] Samples containing one or more suspicious values can be processed following the same methods as those proposed for missing values.

[0053] It is also possible to prepare the missing, outlying or exceptional data in one or more clusters so as to regroup them into the same category for application to them of a particular processing (for example: a weighting, a "false alarm" category, etc.).

[0054] In one variant, one calculates for each input variable its explicative power (or discriminant power) in relation to the phenomenon under study. This process enables, on the one hand, to select the preponderant variables in a list and to eliminate minor variables and, on the other hand, to explain the phenomenon being studied. Preparation of the data can be performed by segmentation algorithms which can, for example, be of the "decision tree" or "support vector machine" type.

[0055] There is preferably associated with each value of a "nominal" variable (for example the zip code or "SIC" code), a table of values expressing its significance in relation to the phenomenon under study (for example: number of inhabitants of the town, income level of the town, average age of the town inhabitants, etc.). It is then feasible to encode the nominal variables using a table of Boolean or real variables.

[0056] For time variables of type $X(t)$, the preparation module also enables automatic generation of columns corresponding to the variable X taken at different anterior or posterior times. Thus, the variable $X(t)$ comes to be substituted by a group of variables: $\{X(t-kdt),...,X(t-dt),X(t),X(t+dt),...,X(t+ndt)\}$.

[0057] In flow modeling applications, time data (date) are transformed by applying transfer rules stemming from the knowledge of the phenomenon under study. For example, for financial flow, when a day is a holiday the associated amounts are transferred according to a set rule in part over the preceding days and in part over the following days with weighting

coefficients.

[0058] It is also feasible to process the flows (for example, financial exchanges) by identifying the periodic payment dates (for example, monthly payment dates) and applying the transfer rules governing each payment date (for example: if the payment date falls on a holiday, transfer the transactions to the following day, etc.).

[0059] The data preparation module offers all of these functionalities on a unitary basis but also allows the user to combine these processings by means of a suitable control language. All these data preparation functionalities are also accessible to the preparation exploration module. The preparation process is terminated preferably by a statistical normalization operation.

[0060] Due to its novel technology, the modeling module (3) associated with the performance analysis module (4) and the optimization module (5), makes it feasible to take into account a large number of model input parameters while controlling accuracy and robustness of the model. It is perfectly suitable for the data preparator described above, which is likely to generate a very large number of often correlated explicative columns, and it makes it feasible to obtain simple, robust and efficient models.

[0061] The modeling module uses a history of the data after preparation. It can be used on all these data, but produces all of its performance when it is only used on part (the rows) of these data, with this part being defined by the optimization module (5) as a function of the results determined by the performance analysis module (4).

[0062] The modeling module proceeds in the following manner:

- the table of the input data after preparation and generation of the Dth order variables constitutes a matrix called [X] of p columns, the outputs corresponding to these inputs constitutes a column vector [Y];

- one constructs a matrix [Z] from the matrix [X] by completing it to the right by a column of 1;

- the model vector [w] is obtained by the following formula:

-14-

$$[w] = (^t[Z]\,[Z] + \lambda\,[H])^{-1}\,(^t[Z]\,[Y])$$

where [H] is a positive defined matrix and λ is a scalar, the introduction of the regularizing term $\lambda[H]$ making it possible to compensate for the ill-conditioning of the matrix $^t[Z]\,[Z]$ by assuring the existence of the inverse matrix $(^t[Z]\,[Z] + \lambda\,[H])^{-1}$.

[0063] The output y* of the model for an input vector [x] = (x1, ..., xp) is obtained by adding a constant equal to 1 at the back of the vector [x], so as to thereby obtain the vector [z] = (x1, ..., xp, 1), then in carrying out the scalar product between the vector [w] and the vector [z], i.e., $y^* = w_1 x_1 + \ldots + w_p x_p + w_{p+1}$.

[0064] There are many possible forms for the matrix [H] used in the construction of the model. In a simple form, the matrix [H] can be written:

$$\begin{bmatrix} 1 & & (0) & \\ & \ddots & & \\ (0) & & 1 & \\ & & & 0 \end{bmatrix}$$

[0065]The matrix [H] can also take on more developed forms, notably:

$$\begin{bmatrix} (1-1/p) & -1/p & \cdots & -1/p & 0 \\ X & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & -1/p & \vdots \\ \vdots & & & \ddots & (1-1/p) & 0 \\ X & \cdots & \cdots & X & 1 \end{bmatrix}$$

[0066] The different forms of the matrix [H] make it feasible to obtain satisfactory models but the user could favor one or the other of the forms according to his *a priori* knowledge

-15-

of the form of the desired model.

[0067] The very construction of the model and of the perturbation matrix [H] confer on the parameter λ or on the vector of parameters $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ particular properties, and notably that of acting on the effective capacity of the learning structure by controlling the trade-off between the model fit and its stability.

[0068] The value of the parameter λ can be set *a priori* by the user according to his knowledge of the phenomenon to be modeled. However, the invention yields its full potential when the selection of the value to be used for the parameter λ is assigned to a module for the automatic determination of the optimal value.

[0069] The purpose of the optimization module (5) is to select the optimal value of the pair of parameters (D, λ) or (D, Λ). For this, the optimization module splits the historical data set after preparation and generation of the Dth order variables into two subsets. The matrix [X] used by the modeling module is only constituted of data from the first subset (referred to as "training space"). The optimization module automatically activates the modeling module while varying the parameter λ. The performances of the models constructed in this manner are evaluated on the untrained data of the second subset (referred to as "generalization space") and the parameter λ selected by the optimization module for the order D considered is the one that permit to obtain an optimum of performances on this second subset according to a validity criterion of the model.

[0070] The data partition can be performed directly by the operator, but it can also be handled by the system in various manners. Thus, the base data partition module can perform a pseudorandom sampling or a sequential sampling to constitute the training and generalization subsets for example: 70% for training and 30% for generalization. In one variant, there is constituted a third subset (referred to as "forecasting space" or "test space") containing data that were used neither for training nor for generalization. Thus, the data partition module can achieve a first sequential sampling into two subsets (for example: the

first 80% of the data for the first subset comprising the training and generalization data, and the remaining 20% for the second subset comprising the test data), then a second pseudorandom sampling on the first subset (for example 70% in training and 30% in generalization for splitting the training and generalization subsets).

[0071] The data partition module can also execute a sampling of the type involving selection of one (or more) sample(s) according to a law that has been programmed in advance (for example: every N samples) for the generation of the training, generalization and/or test subsets.

[0072] The performance analysis module (4) evaluates the performances of the obtained model in relation to certain criteria, the performances being evaluated either on the basis of the training space, i.e., on the data used for the calculation of the matrix [X], or on the generalization space, i.e., on the data that did not participate in the training but were used for the selection of the pair of parameters (D, λ) or (D, Λ), or on the test space, i.e., on the data that were not used for either training or generalization. The performances are evaluated by comparing on the designated space the vector [y], corresponding to the real value of the output, with the vector [y*], corresponding to the value of the output obtained by application of the model. The comparison can be done with classic statistical error indicators, with or without screening. For example, one can use the mean error, or the mean quadratic error or the mean quadratic error weighted by a cost function.

[0073] The model validity criterion used by the optimization module can be selected by the operator from among all of the possibilities offered by the performance analysis module.

[0074] The analysis module also enables filtering the data of a history either in rows or in columns. The row filter criterion relates to the modeling error. This criterion allows separation of the individuals conforming to the model from the nonconforming ones. The nonconforming individuals can be due to anomalies found at the level of the sensors, but they can also reveal an abnormal or original behavior, information which can be very valuable according to the context.

-17-

[0075] The column filtering criterion is performed as a function of the model vector [w]. This enables arranging in order the factors influencing the phenomenon as a function of their positive or negative contribution to the phenomenon.

[0076] The preparation exploration module (6) constitutes the second level of adjustment of the capacity of the learning structure. This module links together the modelings (with or without optimization of the scalar λ) by changing the preparation of the data at each step. This module uses a description of the possible preparations provided by the user. This description defines in an ordered way columns, groups of columns and preparations operating on these columns or groups of columns. For example, the description of the possible preparations can define among the variables of the base data set:

    - a possible polynomial development of column 1, from 1 degree at a minimum to 5 degrees at a maximum;

    - a possible trigonometric development of column 2 from 1 degree at a minimum to 7 degrees at a maximum;

    - a possible multivariable polynomial development on columns 4 to 8 from 1 degree at a minimum to 3 degrees at a maximum;

    - all or part of the other columns without specific processing.

[0077] This description enables formalization of the user knowledge in relation to the phenomenon to be modeled. The preparation explorer thus relieves the user of the tedious tasks of exploration of the possible preparations by carrying out the preparation of the data, the modeling, analysis of performances and recording of the test references and the results obtained.

[0078] This exploration is performed by means of the parameters left free in the description filled out by the user. The explorer can activate different methods in order to achieve this function. Among these methods, the simplest is the systematic exploration of all the possible combinations in the parameters left free by the operator. However, this method can be very costly in terms of calculation time, given that the number of calculations increases

exponentially with the number of parameters.

[0079] Another method consists in executing random sampling in the possible parameters and then sorting the results in a manner so as to approach the zones of greatest interest.

[0080] A third method consists in implementing a control of the capacity of the second level learning process. For this, one uses the fact that for each type of development (polynomial, trigonometric, etc.), the capacity of the learning process increases with the parameter (degree of development). The method starts from a minimal preparation (all of the parameters are at their minimum), and then it envisages all of the possible preparations by incrementing a single parameter. The method launches a modeling for each of the obtained preparations and selects from among the set of obtained models the one that led to the best performance according to a certain criterion.

[0081] Depending on the objective of the user, this criterion can be:

- a minimum of error with or without screening on the unlearned data;

- the ratio between one of the preceding criteria and the capacity of the learning structure after preparation (this capacity can also be approximated by means of known formulas);

- the ratio between the increase in one of the preceding criteria and the increase in the capacity of the learning structure;

- a function increasing with an error criterion such as described above, and decreasing with the capacity of the learning structure.

[0082] The exploitation module (7) enables the tool to transmit the modeling results to a user or to a host system. In a simple version, it can calculate the output of the model evaluated on unlearned data and produce indicators regarding the reliability of the estimation. In a more developed version, the exploitation module can transmit to a host system the generated model, its preparation and its performances. In an even more developed version, the tool is entirely monitored by the host system, such as an industrial process control system, for example, by conferring on it novel potentialities in terms of capacity of adaptation to a

-19-

complex and changing environment.

[0083] A post-processing function (which can be derived from the coefficient λ) allowing calculation of the accuracy of the generated model on new unlearned data can be applied to the result.

[0084] When the database is only composed of few elements characteristic of the phenomenon to be modeled, the training, generalization and forecasting subsets can be not disjoint (for example: use of data belonging to the "training space" subset for generating the "generalization" or "forecasting" spaces).

[0085] The prepared data can be shared among different uses of the data modeling process in accordance with the invention.

[0086] The data set is managed in a specific environment ensuring the availability of the information by using, for example, a file system, a database or a specific tool. It is possible to provide simultaneous access to the data to multiple users. For this purpose, one defines a relational structure containing the variables, the phenomena to be modeled and the models for storing and managing the base data set and the formulas descriptive of the phenomena.